

How not to have millions in inventory and still disappoint the key customer

Stock high and shortages anyway is not bad luck. It is what happens when nobody owns the total. Seven moves to fix it, in order.

Ask an industrial business who decides how much inventory it holds and you get several answers. Service keeps a shelf for the installed base. Installation keeps its own kit, because it was let down last year. Planning buys ahead because suppliers quote sixteen weeks. Finance sets a number at budget time and watches it get missed. Each answer is reasonable. Added together, they are the balance sheet.

Stock is high and the shortages continue, often in the same month. A machine waits at the key customer's site for a board that costs less than the freight to fly it in. Asked who owns the total, the honest answer is nobody. That is a decision-rights problem, and it has a design answer.

1. Separate three jobs

Deciding how much service to give is a commercial decision. It belongs to whoever owns the customer promise: sales for production lead time, service for uptime and response. Calculating the stock to deliver it is arithmetic, and belongs to planning, with one method for every pool. Holding the stock and keeping the record true belongs to a custodian.

When the three jobs collapse into one, each department sets its own quantity and pads for its own risk. The same risk gets covered twice, and the total belongs to no one.

Service level is a commercial decision. Safety stock is a calculation.

2. Cut the list two ways

The familiar cut is 80/20. ITW, where I learned it, taught it on profitability: put the best service behind the customers and products that make the money. In practice nobody knows which customers are profitable. Years of activity-based costing rarely settle it. So start with volume. Volume does not prove a customer is profitable, but it gives you the chance of profit. The top of that cut is the key customer this paper is about. Applied to parts, the same cut on annual spend tells you where to pay attention.

The second cut is runners, repeaters and strangers, on how regularly demand repeats. A runner is needed almost every month, a repeater on a rough cycle, a stranger two or three times a year. Set the breaks where the frequency curve breaks; ten months in twelve and four are typical starting points. The approach came out of Lucas Industries' just-in-time work, which is where I learned it. It tells you how to run an item, not how much attention to give it.

The two cuts disagree, and that is the point. A \$50,000 chamber ordered twice a year is at the top of the spend cut and a stranger: watch it closely, do not stock it, buy to order and quote the real lead time, unless it is on the key customer's

critical list. A \$12 label set used every week is at the bottom of the spend cut and a runner. It stops a build as surely as the chamber. Put it on two bins.

3. Match the planning method to the cut

The long argument between MRP and Lean is mostly an argument about which items you mean.

MRP (material requirements planning, the engine in every ERP) is right for deep bills of material, long uneven lead times and lumpy demand. Its weakness is that it assumes fixed inputs. Planners pad lead times, the plan reaches further onto forecast, and nothing in the logic stops stock piling up. Pull, where a signal such as a kanban card replaces only what was used, is right where demand repeats. Ohno called flow the basic condition, and leveled production the next. On a stranger, a pull loop holds stock that may never be used.

So set the method by segment. Runners replenish by pull, with MRP sizing the loops and placing long-lead commitments. Repeaters run on MRP with calculated safety stock. Strangers are bought to order. Demand-driven MRP is often sold as the shortcut. A review of 57 papers up to 2020 found no quantitative results from real use, and it does not decide your service levels for you.

4. Set service first, then take out stock

The owner of the promise proposes a service level per segment. Planning prices it. The inventory council ratifies it against the cash envelope, and if the two conflict, the conflict goes up with its cost attached, not quietly trimmed. Lean then removes the stock not needed to meet the level. When cash is tight, the key customer's items and service spares are funded first. That is service first, written down.

Say which measure you mean. Cycle service is the chance of not running out between deliveries; fill rate is the share of demand met from the shelf. Fill rate usually reads higher, so a cycle-service target set when the business meant fill rate overstocks.

Two facts keep it honest. The last point of service is expensive: assuming normal demand, going from 95% to 99% cycle service needs about 41% more safety stock, and 99.9% about a third more again. And the cheapest stock is the stock you engineer out. Halve a lead time and the safety stock covering demand variation falls by about 29%; on imported parts, lead-time variability is often the bigger term. A committed supplier lead time does more for the customer than another pallet on the shelf.

5. Gate spares before you size them

Service spares are where the margin is and where the cash sits still. Cohen, Agrawal and Agrawal (2006) found businesses earning 45% of gross profit from the aftermarket on 24% of revenue, with service parts commonly turning once or twice a year.

Stock a spare only if its failure stops or degrades the equipment and it cannot be replaced inside the allowable downtime. If a part can be bought faster than it is needed, it does not belong on the shelf.

Then size by installed base and failure rate, with a Poisson calculation for slow movers. Most businesses lack the failure history, so start from supplier lists, corrected by engineering with a written reason, and record every failure from day one. Supplier lists tend to be conservative.

Keep insurance spares apart: a very low failure rate and a catastrophic cost of downtime. They are a capital decision, justified by downtime cost times probability, and they belong outside turns targets.

6. Stop excess at the forecast buy

Much of the excess is created before the order arrives, in material bought ahead to cover lead time. Buy only against the one forecast signed off in sales and operations planning (S&OP), and only what falls inside the supplier lead time. Cap exposure twice, in months of cover and in dollars. Allocate stock to an order the day it lands, and review anything unallocated on a fixed clock.

With a contract manufacturer (CM), liability is set earlier still. Your exposure is what the CM bought against your authorized forecast, inside lead time, that the cancellation terms will not let it return. The CM's own component suppliers typically allow cancellation 30, 60 or 90 days out; custom and end-of-life parts are usually non-cancelable and non-returnable. Negotiate it in the agreement. Authorize it in S&OP.

For stock already on the shelf, set aging triggers and a disposition route: return, transfer, sell or scrap, decided on a fixed clock. The cash was spent when the parts were bought, whatever the accounting does next. Under US GAAP a write-down is not reversed when demand recovers. Cisco's \$2.2 billion excess inventory charge in 2001 is the public version of the arithmetic.

7. Govern rules, not quantities

Finance sets the total envelope and its split. Planning calculates quantities inside it. An inventory council, co-chaired by supply chain and finance, meets quarterly, ratifies service levels, sets weeks-of-supply and dollar caps, and decides exceptions. It watches three numbers: fill rate, weeks of supply, and excess and obsolete stock as a share of inventory. Nobody on the council sets a part quantity.

Two habits make it stick. Every exception carries an owner and an expiry date; an exception without one is a new policy nobody approved. And when a calculated safety stock breaks a cap, it goes to the council with its supply cause attached, not trimmed until it passes. A cap that hides a sixteen-week lead time only moves the shortage to a later month.

None of this starts with new software. It needs a segmented parts list, a policy per segment, one method, written decision rights and a quarterly review of the same numbers. Do that, and the millions come down while the key customer stops waiting.

That is the work I do.

SOURCES

- Cohen, M.A., Agrawal, N., Agrawal, V. (2006). Winning in the Aftermarket. Harvard Business Review.
- Ohno, T. (1988). Toyota Production System: Beyond Large-Scale Production. Productivity Press.
- Olhager, J. (2010). The role of the customer order decoupling point. Computers in Industry 61(9).
- Azzamouri, A. et al. (2021). Demand Driven MRP: a systematic review and classification. JIEM 14(3).
- Parnaby, J. (1988). A systems approach to the implementation of JIT methodologies in Lucas Industries. IJPR 26(3).
- ECIA (2022). NCNR Product Industry Guideline.
- FASB ASC 330-10-35-14; SEC SAB Topic 5.BB. Cisco Systems (2001), Q3 FY2001 results, SEC Exhibit 99.1.
- Silver, E.A., Pyke, D.F., Thomas, D.J. (2017). Inventory and Production Management in Supply Chains, 4th ed.

WHERE THIS GOES NEXT

The market just turned. What now? →

Committed versus cancelable, and what you owe when demand turns.